Modeling Worldwide Gross Box Office Collection for Films 2012-2021

Rachel Goodridge

Model Motivation

What factors can be used to predict how much a movie will make in theaters worldwide?

Tools and Methods

1. Scrape target (Worldwide Gross) and features from films listed in 2012-2021 on Box Office Mojo
2. Investigate dependent and independent variables
3. Build models using the LinearRegression function from the sklearn package
4. Choose the best model based on scores from cross-validation

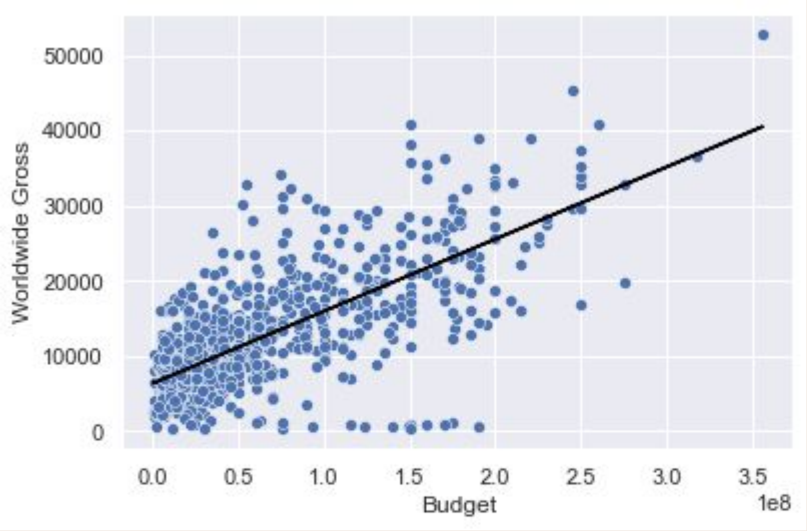
Box Office Mojo
by IMDbPro

Feature Investigation

- Budget
- Running Time
- Number of Theaters
- Distributor
- Genre
- Release Year

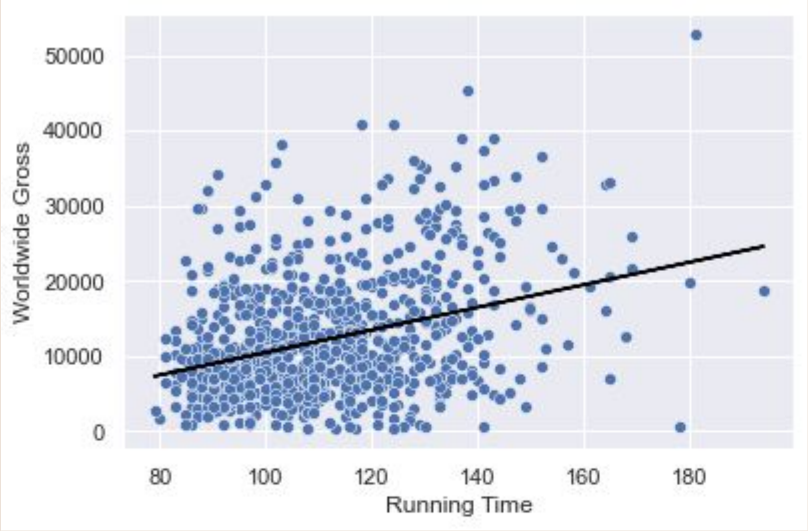
Feature Investigation

- **Budget**
- Running Time
- Number of Theaters
- Distributor
- Genre
- Release Year



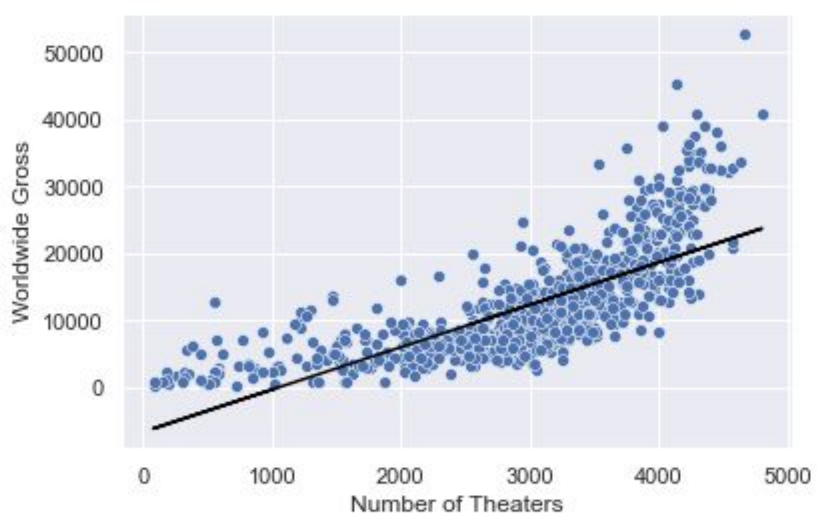
Feature Investigation

- Budget
- **Running Time**
- Number of Theaters
- Distributor
- Genre
- Release Year



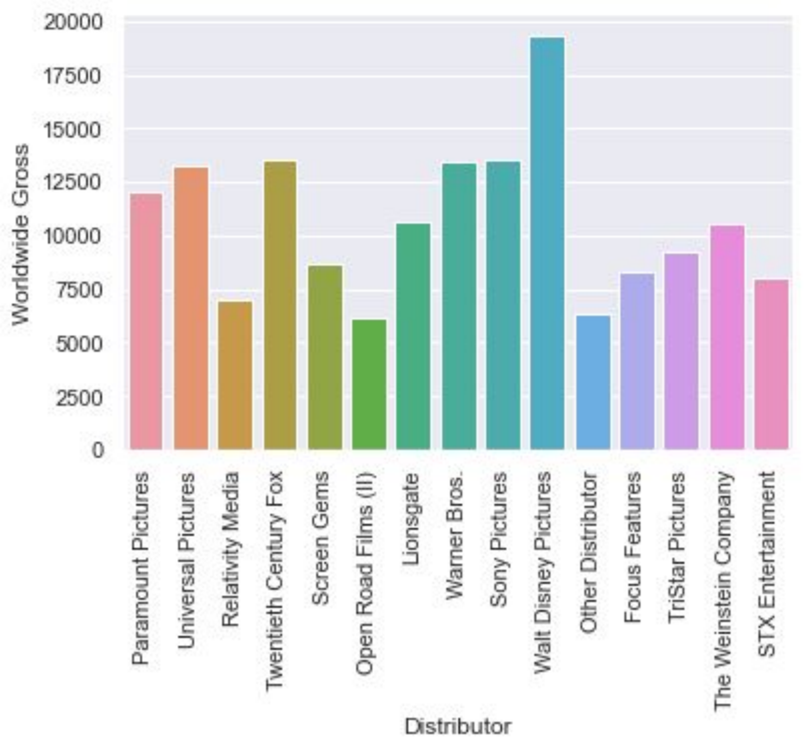
Feature Investigation

- Budget
- Running Time
- **Number of Theaters**
- Distributor
- Genre
- Release Year



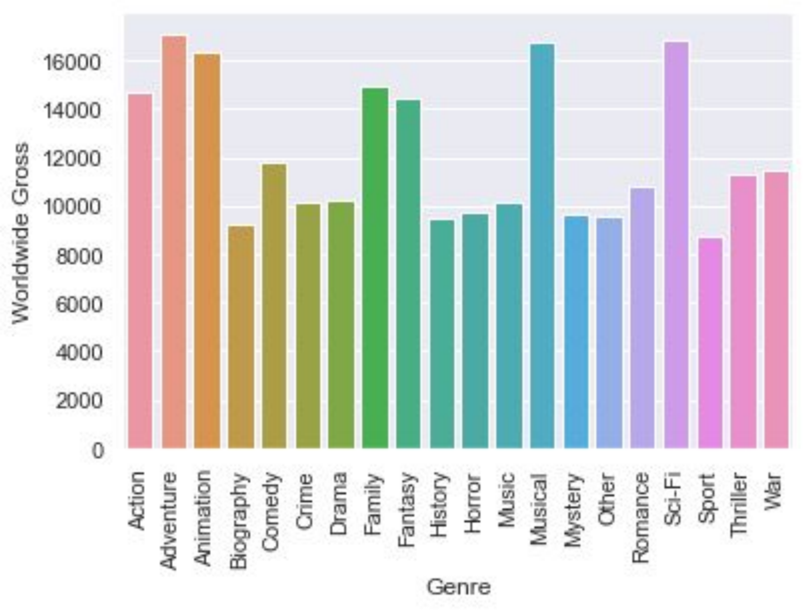
Feature Investigation

- Budget
- Running Time
- Number of Theaters
- **Distributor**
- Genre
- Release Year



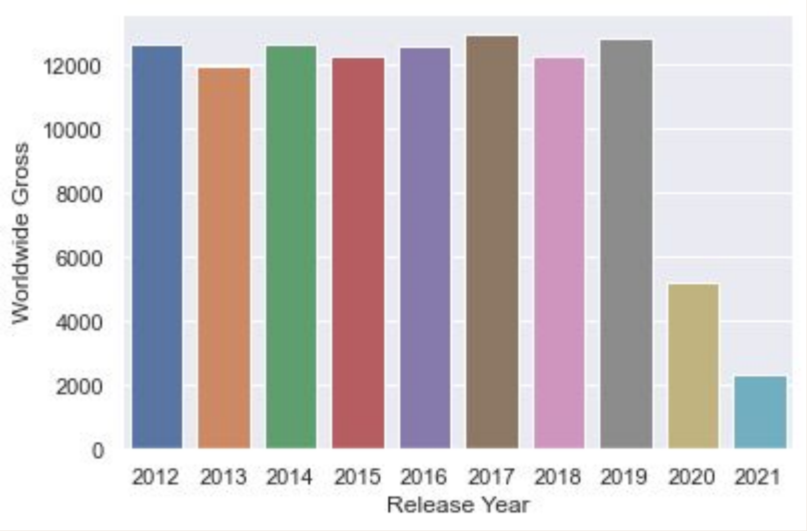
Feature Investigation

- Budget
- Running Time
- Number of Theaters
- Distributor
- **Genre**
- Release Year



Feature Investigation

- Budget
- Running Time
- Number of Theaters
- Distributor
- Genre
- **Release Year**



Multiple Linear Regression Model

$$\sqrt{\mathit{WWGross}} = -7857.75 + (.000056 \cdot \mathit{Budget}) + (40.9309 \cdot \mathit{RunTime}) + (4.01770 \cdot \mathit{NumTheaters}) - (1596.44 \cdot \mathit{OpenRdFilms}) + (1617.38 \cdot \mathit{Universal}) + (2656.12 \cdot \mathit{Disney}) + (3161.74 \cdot \mathit{Animation}) - (1701.24 \cdot \mathit{Family}) - (1233.66 \cdot \mathit{Fantasy}) + (2178.95 \cdot \mathit{Music}) + (1890.15 \cdot \mathit{Musical}) - (2033.96 \cdot \mathit{Sport}) - (4644.61 \cdot \mathit{Year}2020) - (3862.50 \cdot \mathit{Year}2021)$$

Multiple Linear Regression Model

$$\sqrt{WWGross} = -7857.75 + (.000056 \cdot Budget) + (40.9309 \cdot RunTime) + (4.01770 \cdot NumTheaters) - (1596.44 \cdot OpenRdFilms) + (1617.38 \cdot Universal) + (2656.12 \cdot Disney) + (3161.74 \cdot Animation) - (1701.24 \cdot Family) - (1233.66 \cdot Fantasy) + (2178.95 \cdot Music) + (1890.15 \cdot Musical) - (2033.96 \cdot Sport) - (4644.61 \cdot Year2020) - (3862.50 \cdot Year2021)$$

Numerical Data

Multiple Linear Regression Model

$$\sqrt{WWGross} = -7857.75 + (.000056 \cdot Budget) + (40.9309 \cdot RunTime) + (4.01770 \cdot NumTheaters) - (1596.44 \cdot OpenRdFilms) + (1617.38 \cdot Universal) + (2656.12 \cdot Disney) + (3161.74 \cdot Animation) - (1701.24 \cdot Family) - (1233.66 \cdot Fantasy) + (2178.95 \cdot Music) + (1890.15 \cdot Musical) - (2033.96 \cdot Sport) - (4644.61 \cdot Year2020) - (3862.50 \cdot Year2021)$$

Distributors

Multiple Linear Regression Model

$$\sqrt{WWGross} = -7857.75 + (.000056 \cdot Budget) + (40.9309 \cdot RunTime) + (4.01770 \cdot NumTheaters) - (1596.44 \cdot OpenRdFilms) + (1617.38 \cdot Universal) + (2656.12 \cdot Disney) + (3161.74 \cdot \boxed{Animation}) - (1701.24 \cdot \boxed{Family}) - (1233.66 \cdot \boxed{Fantasy}) + (2178.95 \cdot \boxed{Music}) + (1890.15 \cdot \boxed{Musical}) - (2033.96 \cdot \boxed{Sport}) - (4644.61 \cdot Year2020) - (3862.50 \cdot Year2021)$$

Genres

Multiple Linear Regression Model

$$\sqrt{\mathit{WWGross}} = -7857.75 + (.000056 \cdot \mathit{Budget}) + (40.9309 \cdot \mathit{RunTime}) + (4.01770 \cdot \mathit{NumTheaters}) - (1596.44 \cdot \mathit{OpenRdFilms}) + (1617.38 \cdot \mathit{Universal}) + (2656.12 \cdot \mathit{Disney}) + (3161.74 \cdot \mathit{Animation}) - (1701.24 \cdot \mathit{Family}) - (1233.66 \cdot \mathit{Fantasy}) + (2178.95 \cdot \mathit{Music}) + (1890.15 \cdot \mathit{Musical}) - (2033.96 \cdot \mathit{Sport}) - (4644.61 \cdot \mathit{Year2020}) - (3862.50 \cdot \mathit{Year2021})$$

Release Years

Conclusion

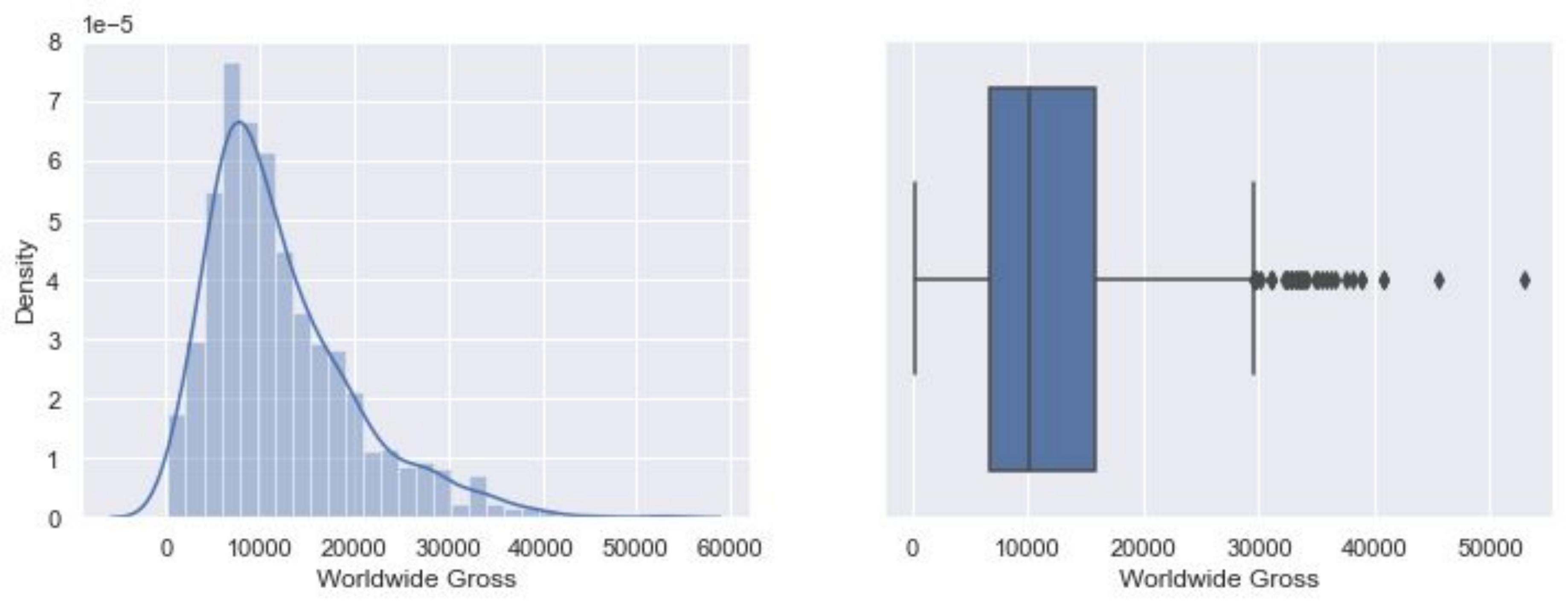
- Films with larger budget, longer running time, and showing in more theaters will likely have a larger Worldwide Gross.
- Universal Pictures and Walt Disney Pictures are more preferable distributors than Open Road Films.
- Films can benefit from including animation, music, and musicals while family, fantasy, and sport genres can be a detriment.
- Recent years have correlated with a decrease in Worldwide Gross (likely due to COVID).

Future Directions

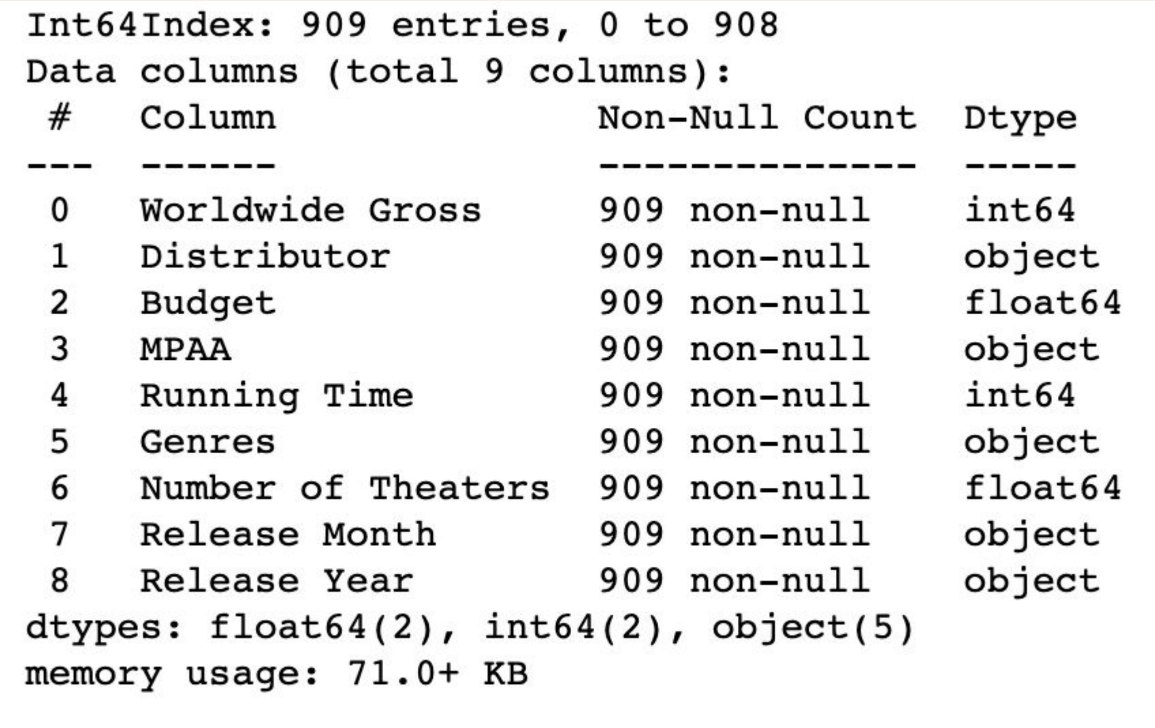
- Search for other important features that are likely missing from this model:
  - Production time, location, cost
  - Actors and directors
  - Pre-existing popularity or familiarity of the story
  - Extent of advertising and social media outreach
- Test out other modeling methods to find a better fit

Thank you!

Appendix

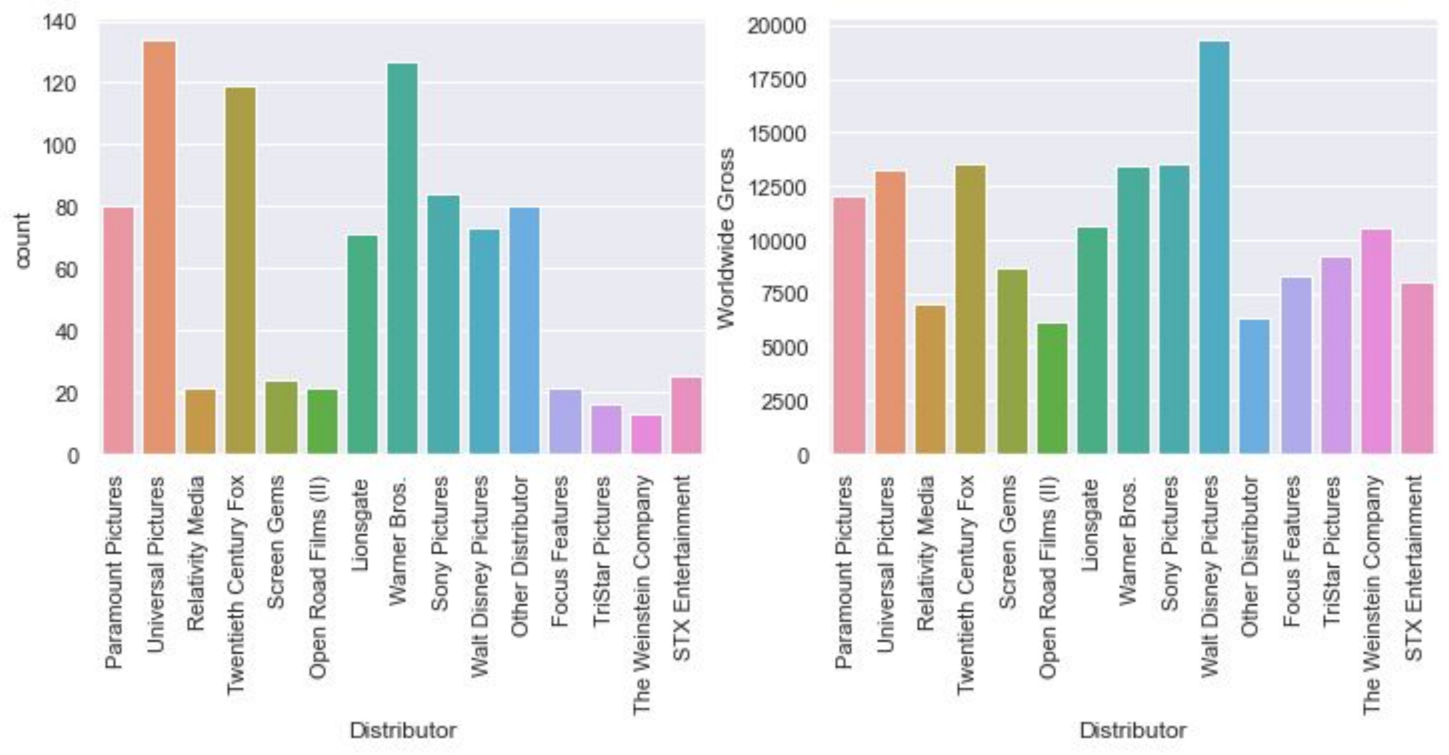

1e−5
8
7
6
5
4
3
2
1
0
Density
0
10000
20000
30000
40000
50000
60000
Worldwide Gross
0
10000
20000
30000
40000
50000
Worldwide Gross

Appendix



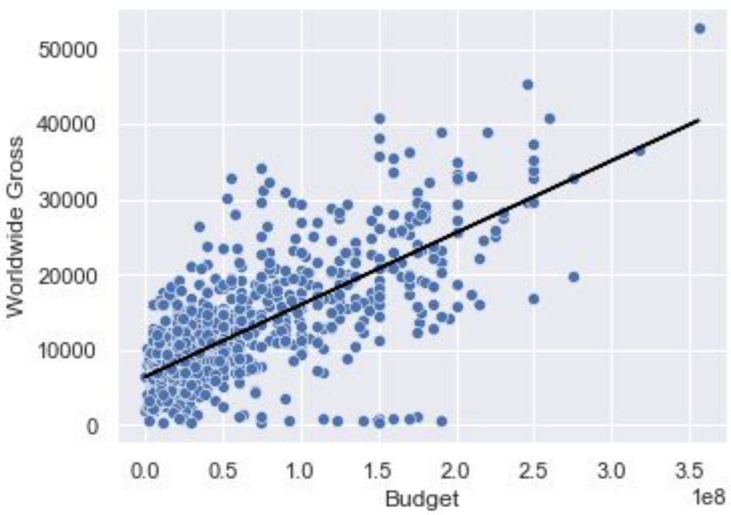
```
Int64Index: 909 entries, 0 to 908
Data columns (total 9 columns):
 #   Column                Non-Null Count  Dtype  
---  ------                --------------  -----  
 0   Worldwide Gross       909 non-null    int64  
 1   Distributor           909 non-null    object 
 2   Budget                909 non-null    float64
 3   MPAA                  909 non-null    object 
 4   Running Time          909 non-null    int64  
 5   Genres                909 non-null    object 
 6   Number of Theaters    909 non-null    float64
 7   Release Month         909 non-null    object 
 8   Release Year          909 non-null    object 
dtypes: float64(2), int64(2), object(5)
memory usage: 71.0+ KB
```

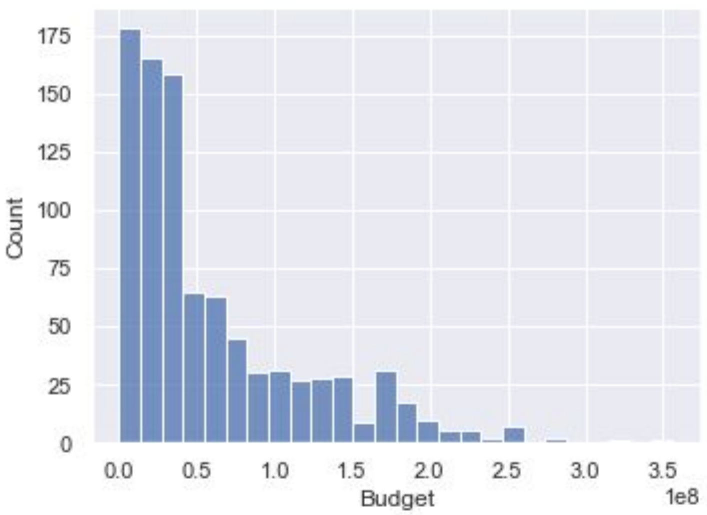
Appendix



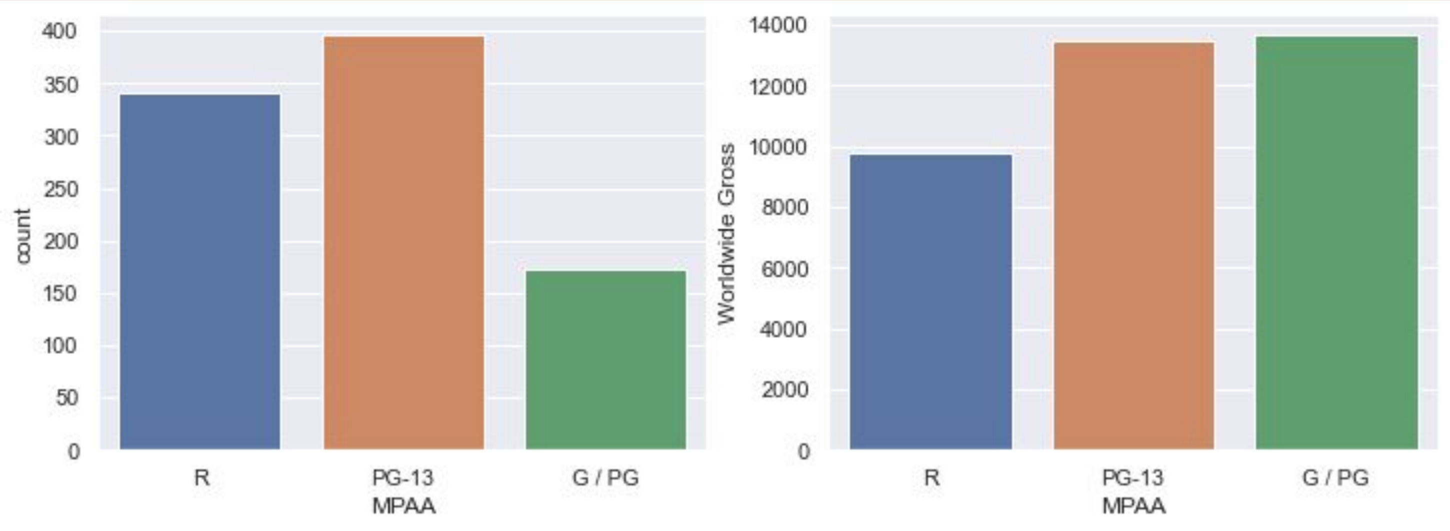
count
140
120
100
80
60
40
20
0
Paramount Pictures
Universal Pictures
Relativity Media
Twentieth Century Fox
Screen Gems
Open Road Films (II)
Lionsgate
Warner Bros.
Sony Pictures
Walt Disney Pictures
Other Distributor
Focus Features
TriStar Pictures
The Weinstein Company
STX Entertainment
Distributor
Worldwide Gross
20000
17500
15000
12500
10000
7500
5000
2500
0
Paramount Pictures
Universal Pictures
Relativity Media
Twentieth Century Fox
Screen Gems
Open Road Films (II)
Lionsgate
Warner Bros.
Sony Pictures
Walt Disney Pictures
Other Distributor
Focus Features
TriStar Pictures
The Weinstein Company
STX Entertainment
Distributor

Appendix

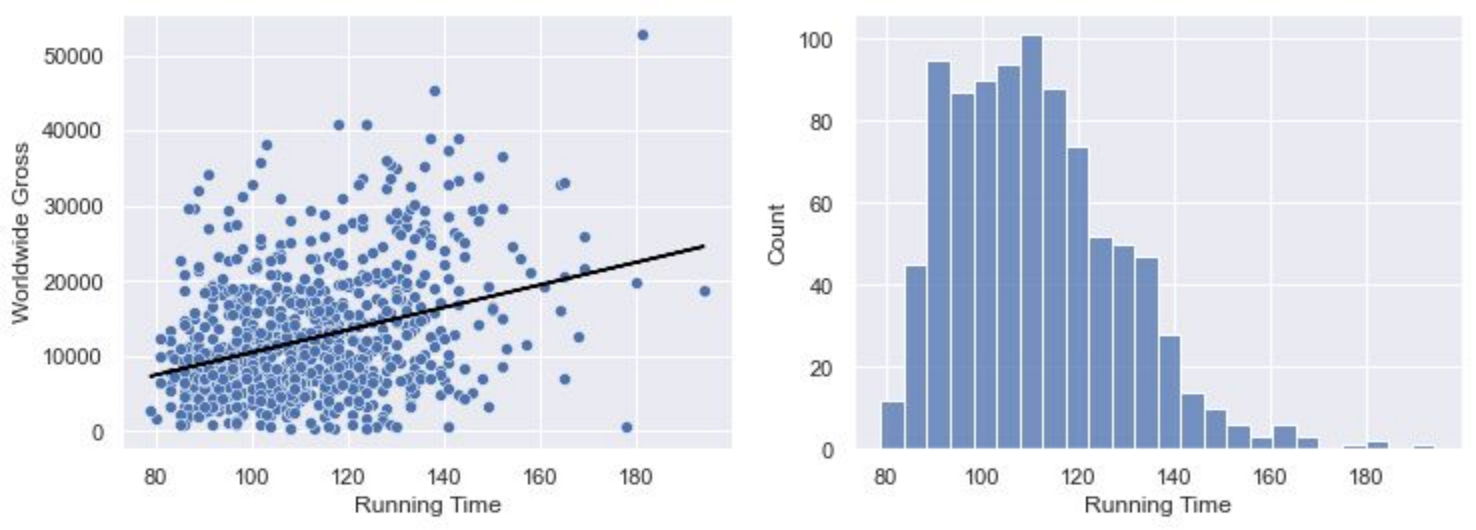

50000
40000
30000
20000
10000
0
Worldwide Gross
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
1e8
Budget


175
150
125
100
75
50
25
0
Count
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
1e8
Budget

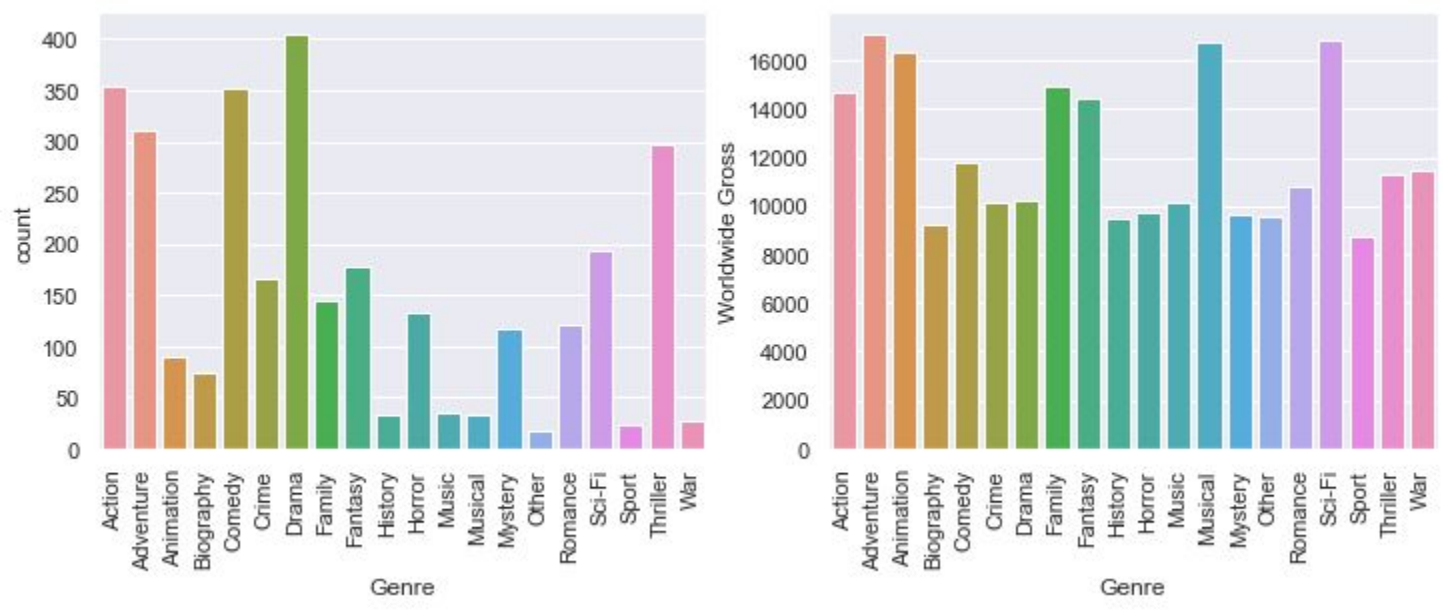
Appendix


count
400
350
300
250
200
150
100
50
0
R
PG-13
G / PG
MPAA
Worldwide Gross
14000
12000
10000
8000
6000
4000
2000
0
R
PG-13
G / PG
MPAA

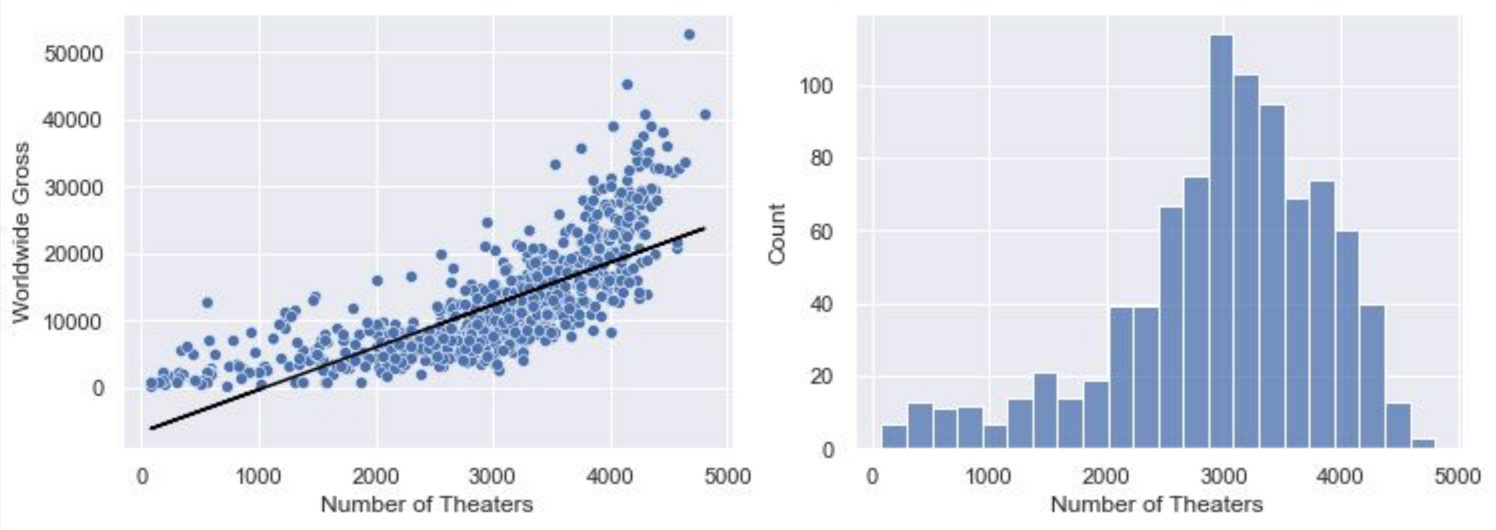
Appendix


Worldwide Gross
50000
40000
30000
20000
10000
0
80
100
120
140
160
180
Running Time
Count
100
80
60
40
20
0
80
100
120
140
160
180
Running Time

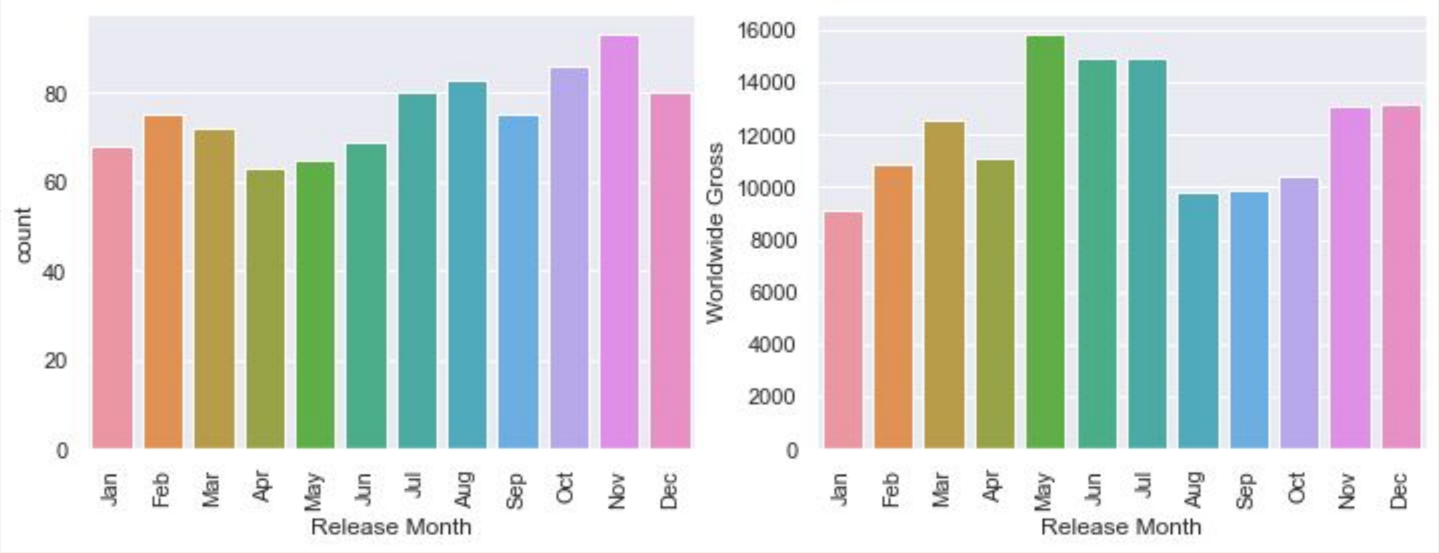
Appendix


count
400
350
300
250
200
150
100
50
0
Action
Adventure
Animation
Biography
Comedy
Crime
Drama
Family
Fantasy
History
Horror
Music
Musical
Mystery
Other
Romance
Sci-Fi
Sport
Thriller
War
Genre
Worldwide Gross
16000
14000
12000
10000
8000
6000
4000
2000
0
Action
Adventure
Animation
Biography
Comedy
Crime
Drama
Family
Fantasy
History
Horror
Music
Musical
Mystery
Other
Romance
Sci-Fi
Sport
Thriller
War
Genre

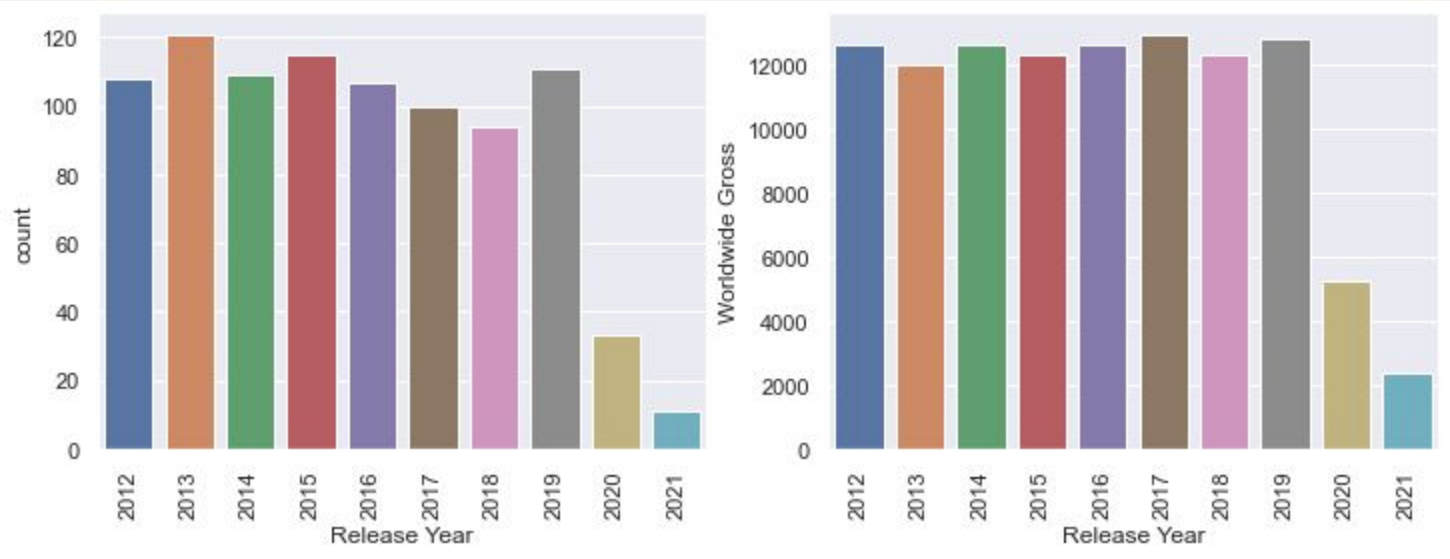
Appendix


50000
40000
30000
20000
10000
0
Worldwide Gross
0
1000
2000
3000
4000
5000
Number of Theaters
100
80
60
40
20
0
Count
0
1000
2000
3000
4000
5000
Number of Theaters

Appendix

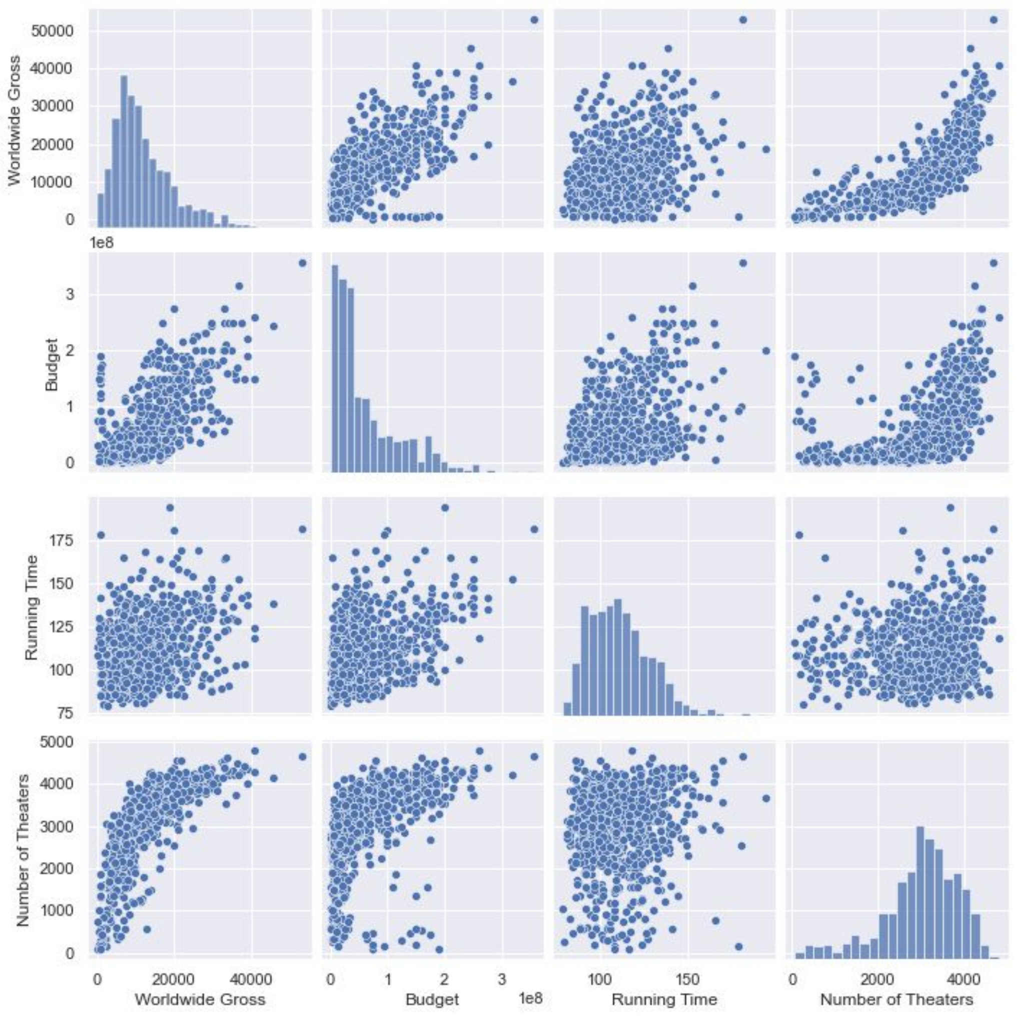

count
0
20
40
60
80
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Release Month
Worldwide Gross
0
2000
4000
6000
8000
10000
12000
14000
16000
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Release Month

Appendix

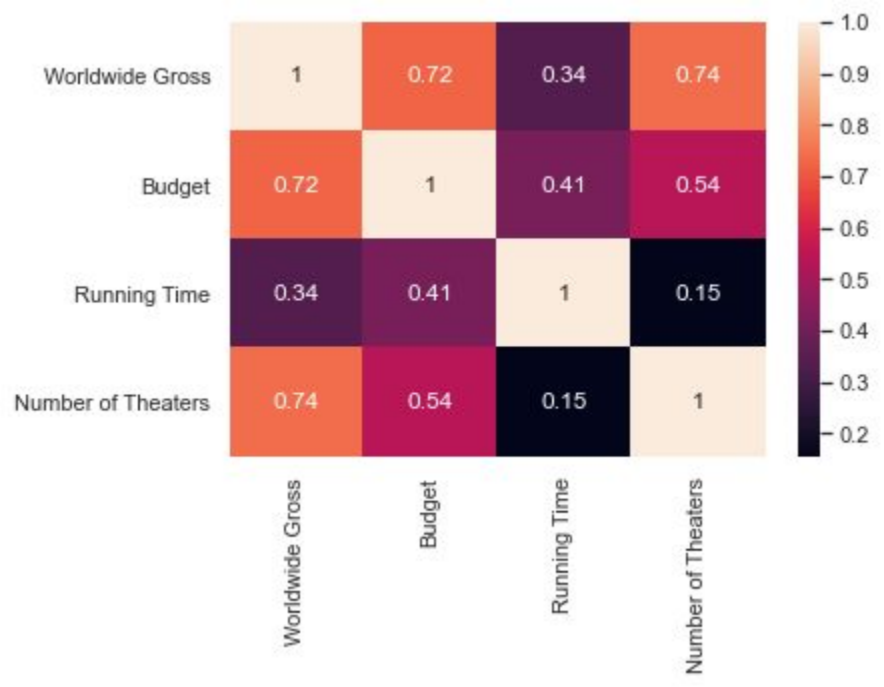


count
120
100
80
60
40
20
0
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
Release Year
Worldwide Gross
12000
10000
8000
6000
4000
2000
0
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
Release Year

Appendix

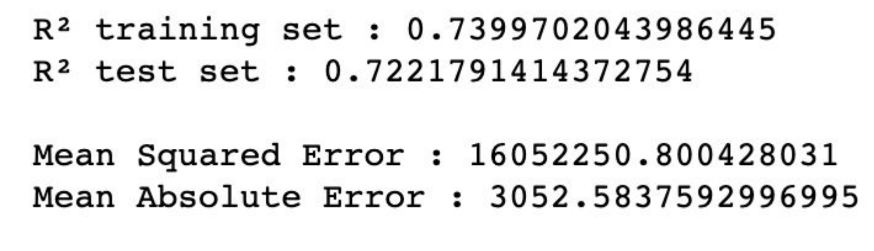

Worldwide Gross
Budget
Running Time
Number of Theaters
50000
40000
30000
20000
10000
0
1e8
3
2
1
0
175
150
125
100
75
5000
4000
3000
2000
1000
0
0
20000
40000
Worldwide Gross
0
1
2
3
1e8
Budget
100
150
Running Time
0
2000
4000
Number of Theaters

Appendix



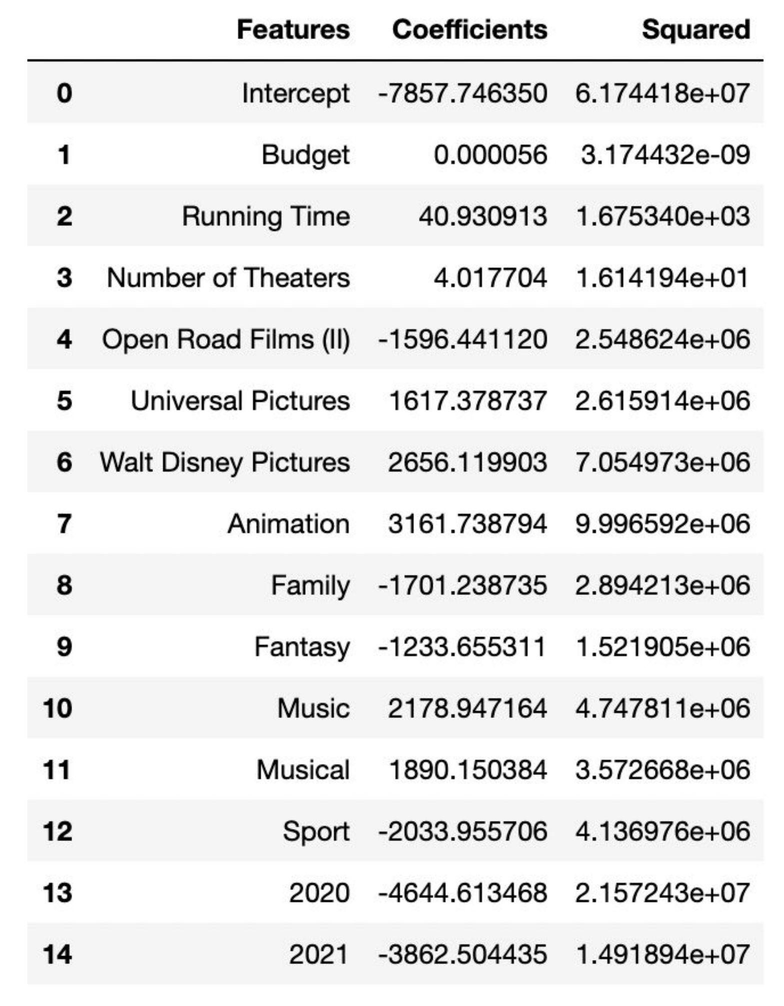
	Worldwide Gross	Budget	Running Time	Number of Theaters
Worldwide Gross	1	0.72	0.34	0.74
Budget	0.72	1	0.41	0.54
Running Time	0.34	0.41	1	0.15
Number of Theaters	0.74	0.54	0.15	1

Appendix



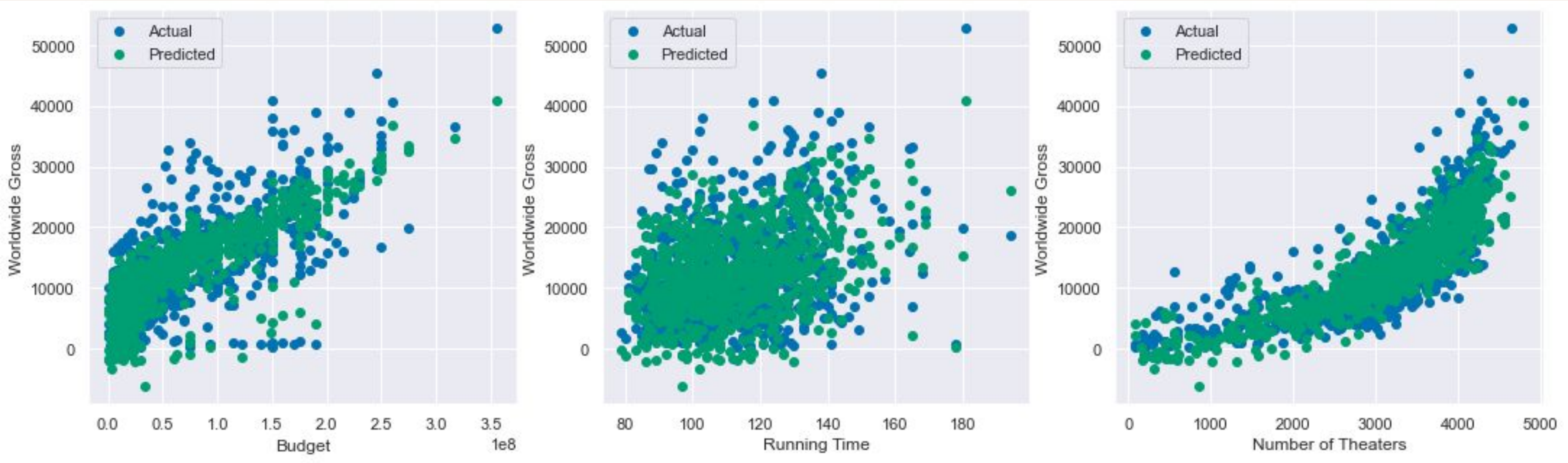
```
R² training set : 0.7399702043986445
R² test set : 0.7221791414372754

Mean Squared Error : 16052250.800428031
Mean Absolute Error : 3052.5837592996995
```



	Features	Coefficients	Squared
0	Intercept	-7857.746350	6.174418e+07
1	Budget	0.000056	3.174432e-09
2	Running Time	40.930913	1.675340e+03
3	Number of Theaters	4.017704	1.614194e+01
4	Open Road Films (II)	-1596.441120	2.548624e+06
5	Universal Pictures	1617.378737	2.615914e+06
6	Walt Disney Pictures	2656.119903	7.054973e+06
7	Animation	3161.738794	9.996592e+06
8	Family	-1701.238735	2.894213e+06
9	Fantasy	-1233.655311	1.521905e+06
10	Music	2178.947164	4.747811e+06
11	Musical	1890.150384	3.572668e+06
12	Sport	-2033.955706	4.136976e+06
13	2020	-4644.613468	2.157243e+07
14	2021	-3862.504435	1.491894e+07

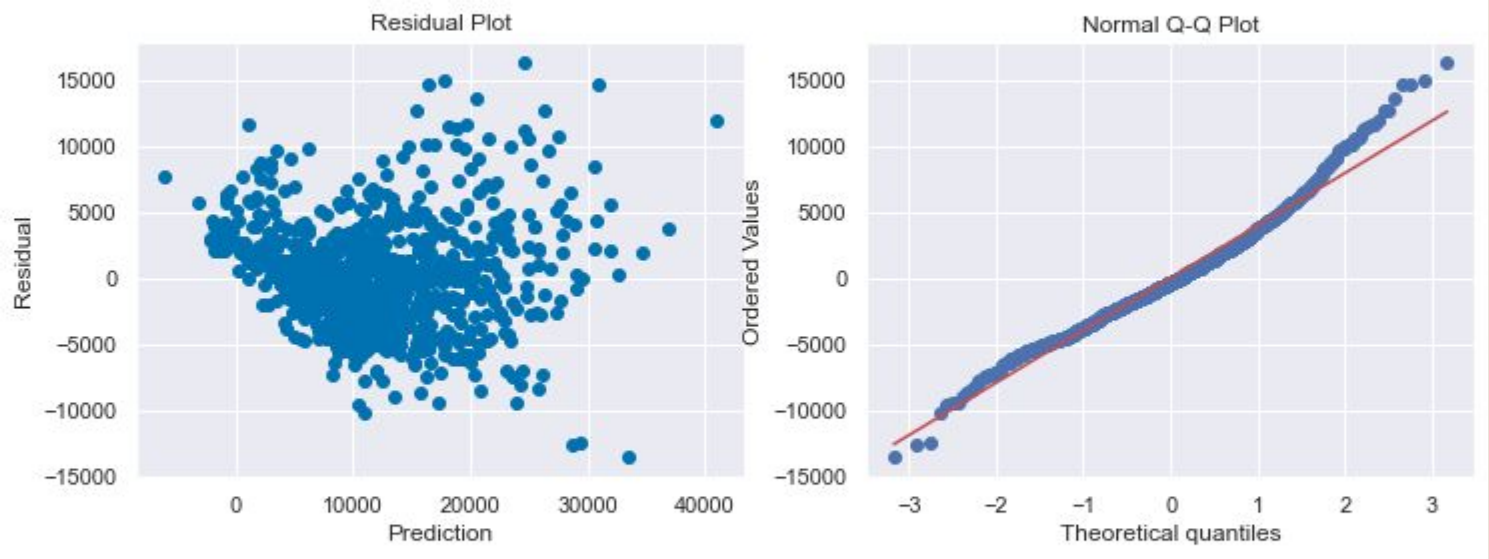
Model Fit on Numerical Data



R^2 score on training set : 0.73997

R^2 score on test data set : 0.72218

Appendix


Residual Plot
Residual
15000
10000
5000
0
−5000
−10000
−15000
0
10000
20000
30000
40000
Prediction
Normal Q-Q Plot
Ordered Values
15000
10000
5000
0
−5000
−10000
−15000
−3
−2
−1
0
1
2
3
Theoretical quantiles